A partial least squares regression approach for long-range intraseasonal and seasonal forecasts

Nat Johnson$^{1,2}$
Michelle L’Heureux$^3$, Stephen Baxter$^{3,4}$, Steven Feldstein$^5$, and Shang-Ping Xie$^2$

39th Annual Climate Diagnostics and Prediction Workshop

$^1$International Pacific Research Center, University of Hawai‘i
$^2$Scripps Institution of Oceanography, University of California, San Diego
$^3$NOAA/NCEP Climate Prediction Center
$^4$University of Maryland
$^5$Penn State University
What is PLS regression?

- A fairly new method (Wold 1966) with limited applications in atmospheric science (e.g., McIntosh et al. 2005, *J. Climate*; Smoliak et al. 2010, *GRL*; Wallace et al. 2012, *PNAS*)

- Sort of a cross between principal component analysis (PCA) and multiple linear regression

- Essentially a multiple linear regression decomposed into steps, where the steps determine “optimal” indices that are used as the predictors in the multiple regression

- These optimal indices are projections onto new variables that are a linear combination of the original predictor variables (latent vectors or PLS components), and each successive PLS component explains less predictand variance than the previous component
1) Calculate correlation coefficients between predictand \( y \) and each gridded predictor time series.

2) Project all predictor maps onto the correlation map to obtain a new predictor time series \( z_1 \).

3) Regress \( y \) on \( z_1 \).

4) Linearly remove \( z_1 \) from \( y \) and all gridded predictor time series, and repeat steps 1-3.
Model complexity and prediction

• Number of PLS components determined through cross-validation
  
  ➢ If the cross-validated residual variance of the first PLS regression exceeds the variance of $y$, then we reject the PLS model
  
  ➢ Otherwise, keep the $n$ PLS components that minimize the cross-validated residual variance

• Double cross-validation: one layer to determine model complexity and one for the prediction
Application to spatial field predictands

\( y_t \): predictand data vector at time \( t \) (e.g., North American T2m)

\( x_1_t \): first predictor data vector at time \( t \) (e.g., tropical SSTs)

\( x_2_t \): second predictor data vector at time \( t \) (e.g., soil moisture)

\( z_1 \) and \( z_2 \): PLS components matrices

\( \mathbf{y}_t \) \( \mathbf{y}_{t,\text{pred}_1} \) \( \mathbf{y}_{t,\text{pred}_2} \)

We want:

\[ y_t = y_{t,\text{pred}_1} + y_{t,\text{pred}_2} + \epsilon_t \]

Component associated with first predictor

Component associated with second predictor

Component residual

EOF analysis

Predicted by some other field of interest (e.g., tropical SSTs)

PLS regression

\[ \mathbf{y}_{t,\text{pred}_1} = \sum_{i=1}^{p} \alpha_{it,\text{pred}_1} \mathbf{e}_i \]

\[ \mathbf{y}_{t,\text{pred}_2} = \sum_{i=1}^{p} \alpha_{it,\text{pred}_2} \mathbf{e}_i \]
This PLS regression approach is conceptually similar to other conventional methods (e.g., CCA and MCA), so what advantages might PLS regression offer?

- Asymmetric: predictand remains distinct from predictor, may aid interpretability
- Can accommodate different sources of predictability for different modes of variability
- Parameter choices arguably more straightforward
- Less prone to overfitting?
Testing the method: Predicting summertime (JJA) T2m over North America (10°-50°N) with May initial conditions

• Predictions of JJA North American T2m anomalies (GHCN CAMS)

• PLS predictions for 1950-2012 with 12 orthogonally rotated EOFs (Varimax)

• PLS predictors: May SSTs (20°S – 60°N, ERSSTv3b) and May North American soil moisture (CPC V2)

• Evaluated against 5 models from the North American Multi-Model Ensemble (NMME) (Kirtman et al. 2014, BAMS): CFSv2 (24 ensemble members), GFDL-CM2.2 (10), CCSM3.0 (6), CMC1-CanCM3 (10), and CMC2-CanCM4 (10)

• Model forecasts are JJA ensemble mean of T2m anomaly forecasts, with a standard calendar month- and lead-dependent bias correction

• Also evaluated against CCA-based forecasts

• Statistical forecasts: Three-year-out (Van den Dool 2009) double cross-validations
How well do the models perform?

Correlations between forecast and verified T2m anomalies 1982-2010
PLS versus CCA
Mean pattern correlations 1950-2012
PLS: 0.46
CCA: 0.40

Performance versus forecast amplitude

Mean RMSE Skill Scores:
PLS: 0.11
CCA: 0.03
Observations and outstanding questions

• Soil moisture appears to add the most skill over the South Central U.S.

Remaining questions

• How much, if any, of the summertime seasonal forecast skill is unrelated to the long-term trend?

• What are the predictor and predictand patterns that contribute to this skill?
Possible intraseasonal application: Bridging the forecast gap in weeks 3-4

**Lead Time**

- 0 ~10 days
- 6-10 Days
- 8-14 Days
- Seasonal
- Monthly
- ~1 month
- ~12 months

**Forecast gap:**

- Based on **initial** conditions
- Based on slowly varying **boundary conditions**

**NOAA CPC products**

?
Preliminary test: Week 3 forecasts of North American winter (DJFM) 2009/10 T2m

**Predictors**

- Previous 7-day extratropical 300 hPa height (z300) + 5-day 200 hPa tropical velocity potential ($\chi_{200}$)
- Previous 7-day T2m anomalies (ERA-Interim)
Preliminary results: Forecast performance

Heidke Skill Scores (HSS)

Mean HSS: 18.9
Mean pattern correlation: 0.36
Mean RMSE skill score: 0.05
Example Week 3 forecast

Week centered on Feb. 15

EOF

z300 and $\chi_{200}$ predictor pattern

Verification

Observed initial conditions
Seasonal mean anomalies

Observed DJFM T2m anoms (°C)

DJFM mean Week 3 forecast T2m anoms (°C)
Conclusions

• Partial least squares (PLS) regression is a useful linear regression approach for univariate predictands with a large set of potential predictors.

• PLS regression may be extended to multivariate predictands, with potential advantages over more well known methods in the climate sciences like CCA and MCA/SVD for some applications.

• A PLS regression approach combined with a rotated PCA appears to outperform dynamical forecast models and CCA for North American summertime T2m forecasts initialized in May.

• A very preliminary analysis suggests that PLS regression may also be useful for prediction and diagnostics on intraseasonal timescales (e.g., predictions for weeks 3-4).

• Future work: addressing sources of predictability in summer, more thorough comparison with alternative approaches, more extensive tests for intraseasonal prediction, extension to probabilistic forecasting.